Data sharing and citation practices – an application of the theory of planned behaviour to social science research practice

Dr. Steve McEachern and Janet McDougall
Australian Data Archive

IASSIST 2014
Presentation Overview

1. About the project
2. Data citation
   a) What we would prefer to see
   b) State of play for ADA
3. Data sharing
   a) A model of data sharing behaviour
   b) Data sharing practices in the US
4. What is still to come?
   a) Australian survey
About the project
Data citation and sharing

- Data citation project
  - ANDS funding
  - Data citation support
  - DOI workflows
- Data sharing project
  - Deposit records
  - User surveys
Australian Survey

- Delayed due to external constraints
- To be conducted in 2014
- This presentation reports on comparative US data
Data sharing
Summary

- Social and economic benefits gained from institutional data sharing are widely recognised by funding agencies, research communities and society.

Despite this only a small number of researchers share data.

- This research will contribute to improving the incidence of data sharing through researching the behaviours of researchers who share data.
Background and Innovation

The aim is to identify who shares data and why

• Acknowledged reasons why researchers
  – do not share data: competition, misrepresentation, cost, ethics, reproducibility.
  – do share data: altruism, data sharing culture, collaboration, esteem, visibility, interest in data issues.

But less is known about researchers who do share data, the focus of this study.

• The objectives are to understand more about researchers who do share data
  – their backgrounds, data management support, experiences and attitudes towards data sharing, and past data sharing behaviour
Approach and methodology

The Theory of Planned Behaviour (TPB) using **direct measures**

TPB is a theory designed to predict and explain human behaviour in specific contexts. The Theory of Planned Behaviour (TPB) is the theoretical model used to approach the problem.

Behaviour for this model:

*As part of the research process, share research data formally for reuse, in the discipline of social science*
Theoretical Framework
TPB Model

Indirect Measures  

- SUBJECTIVE NORM (SN) - socially expected mode of conduct, the perceived social pressure, whether self or important others perform the behaviour.
  - Extent to which researchers should conduct fair exchange for use of other’s shared data.

- ATTITUDE TOWARD THE BEHAVIOUR (AB) – personal positive or negative evaluation of the behaviour.
  - Positive or negative attitude toward availability and misuse of shared data in research

- INTENTION TO PERFORM THE BEHAVIOUR (IB)  
  - Willing to place data in a central repository with no restrictions, to share across a broad group of researchers

Direct Measures

- BEHAVIOURAL BELIEFS  
  - Belief strength x outcome evaluation

- NORMATIVE BELIEFS  
  - Normative belief strength x motivation to comply

- CONTROL BELIEFS  
  - Control belief strength x control belief power

- PERCEIVED BEHAVIOURAL CONTROL (PBC) – control over behaviour and ability to perform behaviour.
  - Control over lack of resources to share research data electronically for reuse

BEHAVIOUR  
- As part of the research process, share research data formally for reuse, in the discipline of social science
Research Questions

1. What are the characteristics of those researchers with intention to perform data sharing behaviour?

2. Are institutional data management procedures associated with intention to perform data sharing behaviour?

3. Is past behaviour associated with intention to perform data sharing behaviour?

4. Is willingness to reuse data associated with intention to perform data sharing behaviour?
• A total of 1329 scientists participated in the survey (2009/10/27 to 2010/07/31) exploring current data sharing practices and perceptions of barriers and enablers of data sharing.

• Survey questions suitable to use for TPB (high reliability).

• Survey precedes data sharing requirements for funding by NSF.

• Demographics:
  – Subset of social scientists (105 cases)
  – Positions (academic and research 77.1%) (graduate students 15.2%) (other 7.6%)
  – Work sector (academic 93.3%) (other 6.7%)
  – Gender ( male 54.1%) ( female 44.9%)
  – Country (Nth America 74.3%) (Europe 15.2%) (other 10.5%)
  – Social science data (survey 74.3%) (other 25.7%)
Statistical method and variables

TPB – Multiple linear regression

• Dependent variable (1-5 strongly disagree-strongly agree)
  - Intention (IB) I would be willing to place all or some of my data in a central repository with no restrictions, to share across a broad group of researchers $\alpha .786$.

• Independent variables (1-2 yes/no, 1-5 strongly disagree-strongly agree, dummies)
  - Attitude (AB) lack of shared data as a major impediment to progress in science and misinterpretation of shared data shared in research due to complexity and poor quality.
    - FA used to highlight responses where agreement with lack of shared data an impediment but not agreement that data to data misuse to overcome respondent complacency (1-5) $\alpha .723$.
  - Subjective Norm (SN) citation, acknowledgement, partial costs retrieval, legal permission, co-authorship (1-2) $\alpha .790$.
  - Perceived Behavioural Control (PBC) lack of funding, standards, time, place to put data (1-2) $\alpha .674$.
  - Organisational support for data management -> formal data management, tools, funds, training processes during and beyond the life of the project (1-5) $\alpha .898$.
  - Share data -> I share my data with others and make it easily accessible (1-5) $\alpha .804$.
  - Reuse data -> I would use other researchers data if it were easily accessible (dummy)
  - Gender M=0 F=1 , Age (dummies).
Hypotheses

+ **H1** Positive attitude (AB) towards data sharing will increase intention to share data.

- **H2** Higher levels of acknowledgement of fair exchange for use of others’ shared data will lead to higher levels of intention to share data, with reference to respondent’s own estimate of conduct of fair exchange - subjective norm (SN).

- **H3** Increased perceived behavioural control (PBC) over resources to share data electronically will increase intention to share data.

+ **H4** Past data sharing behaviour will increase intention to share data.

- **H5** Data management support will lead to higher levels of intention to share data.

+ **H6** Willingness to reuse of others’ shared data will increase intention to share data.
### Variables

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTENTION to share ALL or SOME data with NO restrictions</td>
<td>102</td>
<td>3</td>
<td>15</td>
<td>10.73</td>
<td>2.922</td>
</tr>
<tr>
<td>ATTITUDE factor2</td>
<td>105</td>
<td>-2.3706</td>
<td>2.2050</td>
<td>0.000000</td>
<td>.9951807</td>
</tr>
<tr>
<td>SUBJ NORM</td>
<td>84</td>
<td>11</td>
<td>22</td>
<td>15.13</td>
<td>2.780</td>
</tr>
<tr>
<td>PBC</td>
<td>105</td>
<td>4</td>
<td>8</td>
<td>6.91</td>
<td>1.226</td>
</tr>
<tr>
<td>Organisational data mgt supp</td>
<td>102</td>
<td>7</td>
<td>35</td>
<td>17.99</td>
<td>7.159</td>
</tr>
<tr>
<td>SHARE data</td>
<td>104</td>
<td>2</td>
<td>10</td>
<td>6.29</td>
<td>2.175</td>
</tr>
<tr>
<td>REUSE data</td>
<td>105</td>
<td>0</td>
<td>1</td>
<td>.83</td>
<td>.379</td>
</tr>
<tr>
<td>Gender 0=F 1=M</td>
<td>104</td>
<td>0</td>
<td>1</td>
<td>.54</td>
<td>.501</td>
</tr>
<tr>
<td>age 25-34 DUM</td>
<td>99</td>
<td>0</td>
<td>1</td>
<td>.30</td>
<td>.462</td>
</tr>
<tr>
<td>age 35-44 DUM</td>
<td>99</td>
<td>0</td>
<td>1</td>
<td>.19</td>
<td>.396</td>
</tr>
<tr>
<td>age 45-54 DUM ref var &gt;54</td>
<td>99</td>
<td>0</td>
<td>1</td>
<td>.26</td>
<td>.442</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Analysis

### TPB Model

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj Rsq</td>
<td>0.130</td>
<td>0.174</td>
<td>0.353</td>
<td>0.534</td>
</tr>
<tr>
<td>DW</td>
<td>2.289</td>
<td>2.371</td>
<td>2.196</td>
<td>2.100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>coeff B</th>
<th>sig</th>
<th>coeff B</th>
<th>sig</th>
<th>coeff B</th>
<th>sig</th>
<th>coeff B</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>15.572</td>
<td>0.000</td>
<td>15.837</td>
<td>0.000</td>
<td>11.056</td>
<td>0.000</td>
<td>7.881</td>
<td>0.001</td>
</tr>
<tr>
<td>Attitude</td>
<td>0.650</td>
<td>0.030</td>
<td>0.791</td>
<td>0.008</td>
<td>0.742</td>
<td>0.005</td>
<td>0.540</td>
<td>0.017</td>
</tr>
<tr>
<td>Subjective Norm</td>
<td>-0.054</td>
<td>0.631</td>
<td>-0.051</td>
<td>0.654</td>
<td>0.017</td>
<td>0.862</td>
<td>-0.003</td>
<td>0.977</td>
</tr>
<tr>
<td>Perceived behavioural control</td>
<td>-0.633</td>
<td>0.020</td>
<td>-0.726</td>
<td>0.008</td>
<td>-0.604</td>
<td>0.013</td>
<td>-0.357</td>
<td>0.090</td>
</tr>
<tr>
<td>Organisational data mgt support</td>
<td>0.022</td>
<td>0.631</td>
<td>-0.004</td>
<td>0.931</td>
<td>-0.003</td>
<td>0.937</td>
<td>-0.003</td>
<td>0.937</td>
</tr>
<tr>
<td>I share my data &amp; it's easily accessible</td>
<td>0.563</td>
<td>0.000</td>
<td>0.435</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would use others' datasets if easily accessible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.403</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender (1 = Female)</td>
<td>1.318</td>
<td>0.039</td>
<td>1.380</td>
<td>0.027</td>
<td>0.753</td>
<td>0.183</td>
<td>0.331</td>
<td>0.494</td>
</tr>
<tr>
<td>Age25_34</td>
<td>-0.151</td>
<td>0.857</td>
<td>-0.598</td>
<td>0.485</td>
<td>-0.361</td>
<td>0.634</td>
<td>-0.616</td>
<td>0.342</td>
</tr>
<tr>
<td>Age35_44</td>
<td>-1.092</td>
<td>0.252</td>
<td>-1.448</td>
<td>0.134</td>
<td>-1.018</td>
<td>0.235</td>
<td>-0.975</td>
<td>0.181</td>
</tr>
<tr>
<td>Age45_54 (&gt;54 ref)</td>
<td>-0.591</td>
<td>0.488</td>
<td>-0.676</td>
<td>0.414</td>
<td>-0.643</td>
<td>0.381</td>
<td>-0.622</td>
<td>0.318</td>
</tr>
</tbody>
</table>
Research Questions

1. What are the characteristics of those researchers with intention to perform data sharing behaviour?

• Positive attitude towards data sharing
• No effect of subjective norms among peers (although a weak measure used)
• LOW perceived behavioural control
• More likely to be female (but disappears if controls for data sharing and data reuse)
• No differences by age
Research questions

2. Are institutional data management procedures associated with intention to perform data sharing behaviour?
   NO – Not in this sample

2. Is past behaviour associated with intention to perform data sharing behaviour?
   YES

2. Is willingness to reuse data associated with intention to perform data sharing behaviour?
   YES – Strongest influence in model
Conclusion

Several results from the TPB model and regression were UNEXPECTED

- **Subjective norms**: Lack of acknowledgement, citation and cost recovery often cited for lack of data sharing behaviour
- **Perceived behavioural control**: Previous studies found barriers (i.e. funding, time, place) typically inhibited data sharing behaviour so control of these would expect higher intention to share – rather we found lower.
- **Organisational data support** – data management processes (training, metadata, standards, storage) remove typical barriers.

We have no clear explanations for these!!

Next steps: the Australian survey – a comparison of data citation
Our Acknowledgements

• This project is supported by the Australian National Data Service (ANDS) through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative, as well as through the Australian National University.

• Other thanks go to:
  – Jared Lyle and Amy Pienta, ICPSR
  – Carol Tenopir, University of Tennessee, Knoxville
  – John Kaye, British Library
    • ODIN project - Session C1, Wednesday 3.45, Quirinis
  – Janet McDougall, ADA
  – Masaoka Shiki: Japanese poet, 1867-1902
Thank you very much

For further information:
Web: http://www.ada.edu.au
Email: ada@anu.edu.au

Questions or comments?
Data citation

- State of play for ADA
- What we want to see
What we would prefer to see

Notes

1. The AuSSSA data were obtained from the Australian Social Sciences Data Archive, Australian National University, Canberra.
2. The response categories for the share ownership questions were: 1. No, don't own any shares; 2. Yes, in 1 company; 3. Yes, in 2-5 companies; 4. Yes, in 6-10 companies; 5. Yes, in more than 10 companies.
3. The proportions of respondents in the various income categories are as follows: $0–20,799: 41.6 per cent; $20,800–31,999: 16.6 per cent; $31,200–36,399: 12.9 per cent; $41,600–51,999: 8.8 per cent; $52,000–77,999: 12.3 per cent; and $78,000 plus: 7.8 per cent.
4. Political ideology is measured with the following question: “In politics, people sometimes refer to being on the left or on the right. Where would you place yourself on a scale from 0 to 10 where 0 means the left and 10 means the right?” (emphasis in original). The original scoring of 0 (left) to 10 (right) is retained for the regression analyses.
5. Inglehart (e.g. 1977, p. 28) often uses a question to measure value orientations where respondents are asked to choose their most important and second most important national goals from the following list of options:
   a. Maintain order in the nation.
   b. Give people more say in important government decisions.
   c. Protect freedom of speech.
   d. Maintain a high rate of economic growth.
   Those who choose options 1 and 3, regardless of the order of their choice, are deemed “materialists”. Respondents who choose any combination of options 2 and 4 are classified as “post-materialists”. The remainder comprise the conceptually less interesting but largest category – “mixed” values. A three-point scale is usually constructed from these categories in the manner we employ in these analyses (1 = materialists; 3 = mixed; 3 = postmaterialists).
6. A dummy variable was included to represent the number of respondents in each category.

References


Dallas Hanson lectures in strategic management at the University of Tasmania. He is interested in the intricacies of share ownership and the rationality of business.

Bruce Tranter lectures in sociology at the University of Tasmania. In addition to share ownership, his research interests include environmental sociology, voting behaviour and values research.
Current state of play

• Lack of tradition
  – Depends on your discipline
  – Broadly: low take-up

• Socialisation:
  – Our PIs don’t often “cite”
  – Can we expect others?

• Systemic constraints:
  – Journals lax (or prohibit?)
  – DDI/Nesstar?
Implementing DOIs

• ANDS “Cite My Data”
  – AU DataCite member
  – See “Cite My Data”
    • http://www.ands.org.au/services/cite-my-data.html

• ADA workflows
  – Have now mapped to our process
  – When to mint ID?

• Metadata needs
  – DOI and DDI
  – Nesstar – can’t yet add?